Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predicting Persistent Creatinine Elevation After the First Recorded ICU Furosemide Event in Acute Heart Failure: A Retrospective Cohort Study

Shuting Kong
Jinhua Central Hospital
Yibin Pan
zqwen628@163.com

Jinhua Central Hospital

Research Article

Keywords: acute heart failure, creatinine, furosemide, intensive care, prediction model, worsening renal function

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11474652/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Predicting Persistent Creatinine Elevation After the First Recorded ICU Furosemide Event in Acute Heart Failure: A Retrospective Cohort Study

Article type: Research article

Authors: Shuting Kong; Yibin Pan

Affiliation: Second Ward, Department of Cardiology, Jinhua Central Hospital, No. 365 Renmin East Road, Wucheng District, Jinhua, Zhejiang 321000, China

Corresponding author: Yibin Pan, Second Ward, Department of Cardiology, Jinhua Central Hospital, No. 365 Renmin East Road, Wucheng District, Jinhua, Zhejiang 321000, China. Email: zqwen628@163.com.

ORCID: Shuting Kong, https://orcid.org/0000-0001-7964-2310

Abstract

Background

During decongestion for acute heart failure, an early creatinine rise may resolve or remain elevated. At an intensive care unit (ICU) reassessment, clinicians cannot know from one value whether a first rise will later persist. Existing furosemide-timed models address death or organ support, and broader heart-failure models address acute kidney injury; neither directly answers this question. We examined later confirmation of the first observed rise at 12 and 24 hours after a recorded ICU furosemide event.

Methods

We studied a single-center retrospective Medical Information Mart for Intensive Care IV (MIMIC-IV) cohort. Time zero was the first positive-dose furosemide event in ICU-linked records, not verified initiation of hospital diuretic treatment. Separate ridge logistic models used baseline creatinine, creatinine change, cumulative urine output, and mean arterial pressure. Internal validation assessed discrimination, calibration, and prediction error; additional analyses examined initial-rise status, a simple creatinine rule, and the interval from ICU admission to the recorded event.

Results

Of 784 patients, 481 (61.4%) had a classifiable creatinine course. At 12 and 24 hours, the classifiable risk sets comprised 433 patients/63 events and 383/57; 38.1% and 57.9% of event patients had already shown a first rise. In complete cases, optimism-corrected areas under the receiver operating characteristic curve were 0.783 (bootstrap percentile range 0.721–0.845; 368/52) and 0.840 (0.791–0.900; 357/51). Held-out areas after within-fold predictor imputation were 0.790 and 0.846. The simple rule based only on a rise already recorded had areas of 0.650 and 0.728. Adding cumulative furosemide dose did not consistently improve prediction.

Conclusions

Among patients with enough creatinine follow-up to classify the outcome, routine measurements separated higher from lower risk of later-confirmed persistent worsening renal function. The models may help frame future studies of scheduled renal reassessment, but their performance has not been established for patients without classifiable outcomes or in another hospital.

Keywords: acute heart failure; creatinine; furosemide; intensive care; prediction model; worsening renal function

Background

During treatment for acute heart failure, removing fluid can coincide with a rise in serum creatinine [1–4]. Changes in renal perfusion and concentration during decongestion can alter creatinine without necessarily identifying structural kidney injury; the clinical meaning depends on the accompanying congestion response and underlying kidney function [5–7]. Clinicians therefore need to interpret a new rise in the context of treatment and subsequent measurements, rather than treating every rise as equivalent.

The course of the rise matters. Studies have reported different prognostic associations for creatinine changes accompanied by congestion relief or hemoconcentration and for changes that remain elevated [8–11]. In one large acute-heart-failure cohort, 11.7% had transient and 20.3% had persistent worsening renal

function, using definitions different from ours [12]. Those categories can be assigned after follow-up, but at an ICU reassessment the clinician may see only a first, still unconfirmed rise—or no rise yet. The immediate question is whether the first post-treatment rise will later be confirmed as persistent.

Existing prediction models answer related but different questions. A furosemide-timed ICU model predicted death, renal replacement therapy, and ventilation [13], whereas other heart-failure models predicted broader acute kidney injury [14–18]. Their outcome definitions do not isolate later confirmation of the first observed creatinine rise after treatment. This distinction matters because a patient with an initial rise already visible at 12 or 24 hours is in a different clinical state from a patient who has not yet shown one. Predicting later confirmation could identify whom to monitor more closely; it does not by itself identify harmful kidney injury or determine diuretic dosing.

We therefore anchored time to the first positive-dose furosemide event recorded in ICU-linked data and evaluated four routinely recorded measures at 12 and 24 hours. We asked whether they distinguished patients whose first post-anchor rise would later meet a specified 48–72-hour persistence rule, and whether cumulative recorded dose added predictive information. We also examined the already-risen and not-yet-risen groups separately to clarify what the models predict.

Methods

Study design and data source

This retrospective intensive care unit (ICU) cohort study developed and internally validated prediction models. Each model used information available at a fixed reassessment time after the medication record, known as a landmark. Data came from Medical Information Mart for Intensive Care IV (MIMIC-IV) version 3.1, a deidentified electronic health record database covering admissions to Beth Israel Deaconess Medical Center from 2008 through 2022 [19,20]. Access required PhysioNet credentialing and a data-use agreement. Reporting followed TRIPOD+AI guidance [21], and we considered PROBAST+AI when describing limitations [22]. Ethics information is provided in the Declarations section.

Participants and treatment time origin

The source cohort comprised 1,364 adult ICU stays with a primary diagnosis of acute or acutely decompensated systolic, diastolic, or combined heart failure. We used International Classification of Diseases, Tenth Revision (ICD-10) codes I50.21, I50.23, I50.31, I50.33, I50.41, and I50.43, and corresponding Ninth Revision (ICD-9) codes 428.21, 428.23, 428.31, 428.33, 428.41, and 428.43. Within each hospitalization, only the first ICU stay was considered. We excluded stays coded for end-stage kidney disease or dialysis status.

Time zero was the earliest positive-dose furosemide event (item 221794) linked to the ICU stay in inputevents, after excluding records marked Rewritten. To ensure that inclusion, laboratory trajectories, cumulative dose, and follow-up shared the same time origin, we considered all available events linked to that stay before applying the eligibility rule. The first eligible event had to fall between ICU admission and 24 hours afterward. Stays with an earlier linked event or only zero-dose records were not eligible. We then retained each patient's first eligible stay. Source-event reconciliation identified eight pre-ICU records omitted by the earlier ICU-bounded export; reconciling them reproduced the archived cumulative-dose fields and prompted the cohort and model analyses to be rerun. We retained all non-Rewritten positive-dose status categories, rather than limiting the primary anchor to FinishedRunning rows. A medication-record status does not establish the first diuretic dose anywhere in the hospitalization.

Outcome and landmark eligibility

The original proposed outcome, persistent harmful acute kidney injury (AKI), occurred in only 22 independent patients and failed the prespecified feasibility gate; we did not model it. The retained outcome was persistence of the first observed post-anchor creatinine rise, termed persistent worsening renal function (WRF), not a complete Kidney Disease: Improving Global Outcomes (KDIGO) AKI diagnosis. Creatinine was extracted from 24 hours before to less than 168 hours after time zero. Baseline was the lowest pretreatment value, not a verified outpatient baseline. Post-anchor values were summarized by their maximum in each six-hour bin. The first bin at least 0.3 mg/dL above baseline marked onset. Persistence required the first available bin from 48 to less than 72 hours after onset to

remain at or above this threshold, without an intervening recovery bin. Recovery within the first 48 hours classified that first episode as negative; later recurrent rises were not reconsidered. If no rise was observed, a negative classification required at least one observed bin at or after 48 hours. Missing baseline or insufficient follow-up left the outcome unclassifiable, as detailed in Additional file 1. The definition permits later recovery and does not imply elevation to discharge. Reliable outpatient baseline values, complete weight-normalized urine criteria, and harmonized kidney-replacement-therapy times were unavailable [23,24].

We considered reassessment at 6, 12, 24, 48, and 72 hours after time zero. Before fitting models, we required at least 40 complete-case events for the five-predictor model; only 12 and 24 hours passed. At each landmark, patients had to remain in the ICU, have enough later measurements to classify the outcome, and not already have a confirmed persistent-WRF endpoint. A patient with an initial but still unconfirmed rise remained eligible. The target was therefore later confirmation that the first observed rise persisted, in a mixed group: some patients already had a rise, while others did not. Because confirmation occurred 48–72 hours after onset, the time from reassessment to confirmation differed between patients. Requiring later laboratory data may also have selected which patients entered the analysis.

Predictors

We prespecified four predictors: baseline creatinine (natural log), change in creatinine from baseline, cumulative urine output between time zero and reassessment [log(1 + mL)], and pre-landmark binned non-invasive mean arterial pressure. For creatinine and mean arterial pressure separately, we selected the latest available six-hour bin beginning at or after the anchor and before the landmark: creatinine change used that bin's maximum creatinine minus baseline, whereas mean arterial pressure used the first value within its selected bin. These are binned summaries, not the last raw measurements before reassessment. Adding cumulative furosemide dose [log(1 + mg)] as a fifth predictor tested whether treatment amount provided further information. We chose these measures for clinical relevance and timing in the extract, not for univariable P values. The model is deliberately centered on early recorded renal and blood-pressure response, not a comprehensive account of congestion or kidney disease. Time-appropriate

congestion severity, left ventricular ejection fraction, chronic kidney status, and vasoactive exposure were not sufficiently established for use as point-of-care predictors. Continuous measures remained continuous. We did not use measurements obtained after reassessment, automated feature selection, class balancing, or synthetic observations. Although selected bins began before each landmark, source charting or result-release delays were not verified, so real-time availability of every input was not established.

Statistical analysis

We summarized continuous characteristics as medians and interquartile ranges (IQRs) and compared outcome groups using two-sided Mann–Whitney U tests. Categorical characteristics were reported as counts and percentages. Comparisons used Pearson's χ^2 test, or Fisher's exact test if any expected cell count was <5. Race was compared once across its four categories using a Fisher–Freeman–Halton test with 200,000 Monte Carlo resamples. Each comparison used available observations for that variable. P values in Table 1 and Supplementary Table S1 were exploratory, without adjustment for multiple comparisons, and were not used to choose predictors. Two-sided $P < 0.05$ was used only as a descriptive significance convention. Model variables were determined by clinical relevance, prior evidence, availability at reassessment, and the frozen analysis plan. Additional reviewer-requested comparisons of classifiable versus unclassifiable patients and of recorded hospital outcomes by persistent-WRF status were descriptive and unadjusted; death and discharge timing relative to renal confirmation was not adjudicated.

Separate multivariable ridge logistic regression models were fitted at 12 and 24 hours. Continuous predictors were standardized within each model fit. The inverse penalty parameter was chosen from 13 logarithmically spaced values from 0.001 to 1,000 using stratified five-fold cross-validation and log loss. The primary analysis used complete cases; a random holdout set was not used because it would reduce the information available for both development and evaluation [25]. Internal validation used 500 patient-level bootstrap repetitions. Standardization, penalty tuning, and model fitting were repeated in each bootstrap sample. Mean optimism was subtracted from apparent performance. We report the area under the receiver operating characteristic curve (AUROC), Brier score, calibration intercept, and calibration slope [26].

Calibration intercept and slope were estimated jointly from logit $Pr(Y = 1) = a + b \times logit(p)$, where p is predicted risk; a is not a calibration-in-the-large intercept obtained by fixing b at 1. The reported 2.5th–97.5th percentile ranges describe the corrected replicate estimates; their nominal confidence-interval coverage was not established. Model coefficients were not interpreted as causal effects or screened using their P values; predicted risks and internally evaluated performance were the primary estimands.

Predictor-missingness sensitivity analysis used stochastic iterative imputation with Bayesian ridge conditional models; the outcome was excluded. Twenty completed datasets were generated to examine coefficient direction and tuning stability. For performance estimation, imputation, standardization, penalty tuning, and fitting were repeated within each training fold and then applied to the held-out fold. We used 10-times repeated five-fold nested stratified cross-validation and averaged each patient's 10 held-out predictions. Patient-level bootstrap resampling with 1,000 repetitions provided intervals conditional on the pooled predictions; it did not repeat the full nested procedure. Coefficient ranges across completed datasets were descriptive and were not Rubin-rule confidence intervals. Patients with an unclassifiable outcome were not imputed into the analysis; predictor imputation therefore did not address outcome-ascertainment bias. No inverse-probability correction for outcome ascertainment or competing-risk analysis of death and discharge was performed.

Receiver-operating-characteristic curves were drawn from the same pooled held-out predictions; their AUROCs are distinct from the optimism-corrected complete-case AUROCs. We compared the existing held-out predictions in patients with and without an initial rise in a bin strictly before the landmark; no subgroup model was refitted. As a simple comparator, a binary rule classified patients as positive when a rise of at least 0.3 mg/dL had already been recorded before reassessment. We also described discrimination among patients whose ICU-to-anchor interval was at most six hours, using the same held-out predictions. These subgroup estimates are conditional, exploratory descriptions, not independent validation. We also evaluated the existing held-out base-model predictions after restricting to patients whose selected anchor included a FinishedRunning medication row; no model was refitted. Decision curves used treat-all and treat-none reference strategies from 2% to 30% in 1% increments; for clinical

interpretation only, these represent hypothetical extra renal monitoring for everyone versus no extra monitoring beyond usual care. One thousand patient-level bootstrap repetitions quantified uncertainty conditional on the stored predictions. No actual monitoring intervention or actionable threshold was tested. Analyses used Python 3.14 with pandas, NumPy, scikit-learn, statsmodels, and Matplotlib; the fixed random seed was 20260930. Table 1 comparisons used SciPy. Reviewer-requested descriptive comparisons used SciPy and the existing pooled predictions.

Exploratory analyses used only the reconciled positive-dose, unified-anchor cohort. We compared the 481 patients with a classifiable outcome against the 303 without one using recorded demographic, diagnosis, severity, treatment, and length-of-stay fields; fields not fixed before the anchor were not treated as baseline predictors. Among classifiable patients, we compared recorded in-hospital death with Fisher's exact test and hospital length of stay with the Mann–Whitney U test by persistent-WRF status. These comparisons were unadjusted, were not used for predictor selection, and cannot establish temporal or causal relationships.

External-data transportability gate

We examined whether the eICU Collaborative Research Database could support an independent evaluation [27,28]. We used the same broad cohort criteria, reassessment times, and persistent-WRF definition. However, its medication extract contained furosemide order times rather than confirmed administration times and did not provide cumulative administered dose. Before estimating external performance, our frozen feasibility rule required at least 40 future events at each reassessment time.

Use of large language models

OpenAI ChatGPT and Codex assisted with language editing, manuscript organization, drafting and review of analysis code, and preparation of tables and figures. WorkBuddy was used for an additional manuscript-consistency review. Numerical estimates were calculated by Python programs from the authorized local datasets and archived model predictions; they were not supplied directly by language-model output. None of these tools is listed as an author. The authors remain responsible for checking the code, analyses, and manuscript against the source data before submission.

Results

Cohort and landmark populations

Of 1,364 source ICU stays, 1,042 had a furosemide record and 1,041 had a positive recorded dose. In 906 stays, the earliest eligible event occurred within 24 hours after ICU admission. Keeping each patient's first eligible stay left 784 independent patients. Persistent WRF could be classified in 481 (61.4%), including 64 events. It could not be classified in 303: 227 lacked pretreatment baseline creatinine, three lacked any later creatinine, 39 had too little follow-up to rule out an initial rise, and 34 had an initial rise but insufficient data 48–72 hours afterward to determine persistence (Fig. 1; Additional file 1: Table S6). Compared with unclassifiable patients, those with a classifiable outcome were younger (median 72 vs 78 years), stayed longer in the ICU (68.3 vs 44.2 hours), and more often had a recorded norepinephrine exposure (59/481 vs 21/303); hospital death occurred in 30/481 (6.2%) versus 35/303 (11.6%) (Additional file 1: Table S9). Diagnostic and first-day severity fields in that comparison can include post-anchor information and were used only to describe selection. These unadjusted differences cannot identify the direction of bias in model performance.

The analysis populations narrowed further at each reassessment. At 12 hours, 695 patients remained in the ICU; 262 lacked a classifiable outcome, leaving 433 patients/63 later-confirmed events (complete predictors: 368/52). At 24 hours, 587 remained; 204 lacked a classifiable outcome, leaving 383/57 (complete predictors: 357/51). Before reassessment, a first rise had appeared in 54 and 73 patients,

including 24/63 (38.1%) and 33/57 (57.9%) of those who later met the persistence definition. Among the remaining event patients, five and seven first rose in the six-hour bin beginning at the respective landmark, and 34 and 17 first rose later. Median time from reassessment to confirmation was 60 hours (IQR 45–75) and 42 hours (IQR 30–60). The models therefore combine later confirmation of an existing rise with prediction of a future rise. Bin-level timing cannot prove that a result in the landmark-start bin was available at the exact reassessment time (Additional file 1: Table S7).

At 12 hours, the risk set had a median age of 72 years (IQR 62–82), and 183 patients (42.3%) were women. Patients with later-confirmed persistent WRF had larger creatinine increases and lower cumulative urine output. Exploratory comparisons also showed differences in first-day Sequential Organ Failure Assessment (SOFA) score and documented renal disease. These P values were not used to select model variables. Findings at 24 hours followed a similar descriptive pattern (Table 1; Additional file 1: Table S1).

The existing held-out base-model AUROC was 0.835 in 54 patients/24 events with a rise before 12 hours and 0.722 in 379/39 without one. At 24 hours, these values were 0.835 in 73/33 and 0.734 in 310/24. A simple binary rule based only on a rise already recorded had sensitivity/specificity of 38.1%/91.9% at 12 hours and 57.9%/87.7% at 24 hours, with binary AUROCs of 0.650 and 0.728 (Additional file 1: Table S8). These exploratory comparisons use the original held-out predictions and do not establish subgroup calibration or incremental clinical benefit. Among patients whose anchor fell within six hours of ICU entry, AUROC from those predictions was 0.775 in 266/38 at 12 hours and 0.850 in 233/35 at 24 hours; no model was refitted for this subset (Table S8). With restriction to an anchor containing a FinishedRunning row, the base-model AUROC was 0.798 in 399 patients/60 events at 12 hours and 0.852 in 353/54 at 24 hours, using the same held-out predictions (Table S8).

Among all 784 patients, the recorded anchor occurred a median 4.1 hours (IQR 2.2–8.3) after ICU entry and 8.3 hours (IQR 4.1–22.0) after hospital admission; 510 (65.1%) anchors were within six hours of ICU entry (Additional file 1: Table S9). At the selected anchor time, 51/784 patients had no FinishedRunning row: 41 had a ChangeDose/Rate row, nine a Stopped row, and one a Paused row. In the 481 patients with

a classifiable outcome, hospital death was recorded in 14/64 (21.9%) with persistent WRF and 16/417 (3.8%) without it; median hospital stay was 282.3 versus 212.2 hours (Table S10). The association is descriptive: testing and length of stay affect who can satisfy the outcome, and death was not ordered relative to renal confirmation.

Model performance

In complete cases, the four-predictor model had an optimism-corrected AUROC of 0.783 (0.721–0.845) at 12 hours (368 patients/52 events). Its Brier score was 0.100 (0.077–0.122), calibration intercept 0.035 (-0.454–0.684), and calibration slope 1.020 (0.760–1.364). At 24 hours (357/51), the corresponding values were AUROC 0.840 (0.791–0.900), Brier score 0.092 (0.070–0.113), intercept 0.085 (-0.521–0.867), and slope 1.050 (0.748–1.474) (Table 2). Parentheses give bootstrap 2.5th–97.5th percentile ranges. The larger 24-hour AUROC does not show that waiting improves care or prediction for the same patients: the risk sets differed. Creatinine change had the largest standardized coefficient; greater cumulative urine output was associated with lower predicted risk.

Adding cumulative furosemide dose changed the corrected AUROC by 0.008 at 12 hours and 0.003 at 24 hours. Brier scores increased by +0.001 and +0.002, respectively; lower scores indicate less prediction error. We assessed these changes together with coefficient stability and held-out performance. Small AUROC differences alone were not treated as evidence that the model would be useful in clinical practice.

Missing-data sensitivity analysis and decision curves

At 12 hours, 40 patients (9.2%) lacked a creatinine-change value and 31 (7.2%) lacked mean arterial pressure. At 24 hours, the corresponding counts were three (0.8%) and 23 (6.0%); other prespecified predictors were complete (Additional file 1: Table S2). With imputation within training folds, held-out AUROC was 0.790 (conditional bootstrap percentile range 0.731–0.848) in 433 patients/63 events and 0.846 (0.792–0.899) in 383/57 (Fig. 4). Brier scores were 0.099 and 0.091, and calibration slopes were 1.171 and 1.039 (Additional file 1: Table S3; Fig. 2). The 12-hour slope point estimate exceeded its

complete-case counterpart of 1.020. A slope above 1 is consistent with predicted log-odds that are less dispersed than the observed outcome gradient; however, the conditional bootstrap percentile range for the imputed-predictor slope was 0.913–1.550 and included 1. The highest base-model predicted-risk decile contained only 44 patients at 12 hours and 39 at 24 hours, contributing to uncertainty toward the right of the calibration plots. These analyses still excluded patients whose outcome could not be classified; percentile ranges were conditional on the saved held-out predictions, not conventional confidence intervals from a fully rerun model-building procedure.

Across thresholds of 2%–30%, base-model net-benefit point estimates were above both reference curves. At thresholds of 10%, 20%, and 30%, net benefit per 100 patients was 7.75, 4.50, 3.43 at 12 hours and 9.08, 6.85, 5.45 at 24 hours (Additional file 1: Table S4; Fig. 3). Here, treat-all would mean hypothetically offering extra renal monitoring to every eligible patient and treat-none would mean offering no extra monitoring beyond usual care. Neither strategy nor a risk-triggered intervention was tested. These internal exploratory calculations do not establish that using the model improves patient outcomes.

External-data feasibility

The extracted eICU heart-failure cohort contained 229 admissions. Only 32 had a usable furosemide order time within 24 hours of ICU admission, and each landmark had just three future events (Additional file 1: Table S5). Neither reached the prespecified 40-event threshold. We did not estimate external performance; this was a feasibility check, not external validation.

Discussion

Principal findings

Four routine measurements distinguished higher from lower risk of later-confirmed persistent WRF among patients whose creatinine course could be classified. The 24-hour model had a numerically higher AUROC than the 12-hour model, but the risk sets differed; this is not evidence that waiting improves

care. The outcome combined a first rise already seen at reassessment with a first rise that appeared later. In the existing held-out predictions, AUROC was 0.835 versus 0.722 in the already-risen and not-yet-risen 12-hour groups, and 0.835 versus 0.734 at 24 hours (Additional file 1: Table S8). These small-subgroup descriptions are not separate model validations. Especially at 24 hours, discrimination partly reflects recognition of an existing change rather than anticipation of a wholly new event. A later recurrence after recovery of the first episode was not counted as positive.

The measured renal response and blood pressure carried more predictive information than total recorded furosemide dose in this dataset. Adding dose produced little stable improvement after the four prespecified measures were included. This finding concerns incremental prediction only. It does not establish that diuretic exposure is unrelated to the renal outcome, or whether furosemide benefits or harms the kidney: dose is adjusted in response to congestion, blood pressure, renal function, and prior treatment response.

Relation to previous work and study contribution

Kamio et al. used repeated furosemide administrations as prediction times for death and organ-support outcomes in Japanese ICUs [13]. Our contribution is therefore the specific question rather than the use of medication timing: after the first recorded ICU event, can measurements at 12 or 24 hours identify patients whose first observed post-anchor creatinine rise will later be confirmed to persist? Other heart-failure models address broader acute kidney injury outcomes [14–18]. Their AUROCs cannot be directly compared with ours because the populations, outcomes, prediction times, and validation methods differ.

Persistent creatinine elevation deserves attention because previous acute-heart-failure studies have found that its prognostic association differs from that of a transient rise [8–12]. Among our 481 classifiable patients, hospital death was recorded in 14/64 (21.9%) with persistent WRF and 16/417 (3.8%) without; median hospital stay was 282.3 versus 212.2 hours (Additional file 1: Table S10). These are unadjusted co-occurrences, subject to selective testing and differences in time available to observe a persistent rise. We did not establish whether death followed renal confirmation or whether this specific endpoint

independently predicted harm. Our endpoint was persistence of the first observed rise at 48–72 hours after onset, not persistence to discharge or a complete KDIGO AKI diagnosis.

Strengths and clinical implications

The study used several safeguards to make internal performance easier to judge. We retained one stay per patient and set a minimum event count before model fitting. Preprocessing and tuning were repeated within resampling. We reported calibration and overall prediction error as well as AUROC. When predictors were missing, imputation was fitted only in training folds, and ROC and decision curves used held-out predictions. Nevertheless, the complete-case models contained only 52 and 51 events. Penalization and bootstrap correction cannot eliminate coefficient instability or make calibration precise in such small samples; the reported percentile ranges should be read with that uncertainty in mind. These safeguards also do not remove selection caused by incomplete later testing, establish real-time availability of every measurement, or replace evaluation in another hospital.

A plausible future use of a high-risk result is to prompt a scheduled repeat creatinine test and a review of congestion, blood pressure, and urine output, rather than an automatic change in diuretic dose. In the decision curves, treat-all and treat-none correspond only to hypothetical extra monitoring for everyone or no extra monitoring beyond usual care. Because neither a threshold nor the monitoring pathway was specified or tested, the curves cannot show clinical benefit. Future studies should verify complete hospital medication history, address patients without enough creatinine follow-up, fix the outcome window after reassessment, and test the unchanged model in an independent cohort with enough events. A causal question about diuretic dosing would require a different design that accounts for time-varying confounding.

Limitations

The principal threat to validity is outcome ascertainment. A WRF outcome could be classified in only 481 of 784 patients (61.4%). Among the 303 others, 227 lacked a pretreatment baseline and 76 lacked sufficient later creatinine data. The groups differed in age, length of stay, severity proxies, and observed

hospital death (Additional file 1: Table S9). Selective ordering and availability of creatinine tests can change which renal trajectories are visible; conditioning development and validation on classifiability may alter the event mix, apparent discrimination, and calibration in either direction. These descriptive comparisons cannot identify the magnitude or direction of bias. Imputing outcomes or weighting the analyzed sample would require assumptions about unobserved creatinine courses and the testing process that these data cannot verify; we did not prespecify or test such scenarios. We did not model death and discharge as competing events. Accordingly, the AUROCs, calibration, and decision curves apply only to the classifiable subgroup, not to all patients present at reassessment.

The lowest creatinine in the 24 hours before the ICU record is not a verified outpatient baseline. Our outcome measured elevation 48–72 hours after the first rise, not persistence to discharge; its timing after reassessment varied. The algorithm did not reconsider recurrent episodes after recovery of the first rise. A six-hour-bin maximum above the threshold does not show that every measurement in that bin remained elevated. Follow-up was truncated before 168 hours after the anchor, limiting confirmation of late rises; patients without an observed rise could be classified negative after a measurement at or after 48 hours without complete 168-hour follow-up. Risk sets mixed patients with and without an initial rise.

Recent weight was often unavailable, so urine output could not be weight-standardized. Ejection fraction was unavailable in the analyzed wide extract, and N-terminal pro-B-type natriuretic peptide was too incomplete for subgroup analysis. Admission-diagnosis kidney-disease flags, first-day SOFA, and a wide-window vasopressor indicator could contain information recorded after the prediction time; adding them to a point-of-care model without temporal reconstruction would risk leakage. Thus the selected predictors mainly describe early observed renal and blood-pressure response, not all determinants of renal dysfunction. The two extraction branches were reconciled against exported source events, and the cohort and analyses were rerun using one positive-dose time anchor. The anchor occurred a median 4.1 hours after ICU entry and 8.3 hours after hospital admission; in 510/784 patients it occurred within six hours of ICU entry (Additional file 1: Table S9). Earlier hospital doses outside these ICU-linked records were not comprehensively assessed. The 51 anchors without a FinishedRunning row and the status-restricted

analysis underscore that this record time is not verified first administration. Input freshness and charting delay were not verified.

Predictor imputation gave similar results among patients with classifiable outcomes but did not address missing outcomes. The 12-hour imputed-predictor calibration slope was higher than its complete-case counterpart, so calibration was not identical across analyses. Intervals around pooled cross-validation predictions were conditional on stored predictions rather than a rerun of the complete nested procedure. We did not develop a separate model for patients without a rise in a bin strictly before reassessment: this group contained 39 events at 12 hours and 24 at 24 hours, so its complete-case event count necessarily remained below the prespecified 40-event development threshold. Exploratory subgroup AUROCs and the short ICU-to-anchor subset estimate came from the existing predictions and may be unstable (Additional file 1: Table S8). Cumulative dose did not capture dose escalation or why clinicians changed treatment. In eICU, order time was not confirmed administration time and too few events were available for external evaluation; no external validation was performed. Practice differences in medication recording, laboratory testing, and ICU admission may further limit transportability.

Conclusions

After the first recorded ICU-linked furosemide event, four routine measurements distinguished higher from lower risk of later-confirmed persistent WRF in the subgroup with classifiable creatinine courses. The distinction between an already observed rise and one not yet seen changes how the prediction should be interpreted. Prospective testing of a defined renal-monitoring action and independent validation are needed before clinical use.

List of abbreviations

AKI, acute kidney injury; AUROC, area under the receiver operating characteristic curve; ICD, International Classification of Diseases; ICU, intensive care unit; IQR, interquartile range; KDIGO,

Kidney Disease: Improving Global Outcomes; MAP, mean arterial pressure; ROC, receiver operating characteristic; SOFA, Sequential Organ Failure Assessment; WRF, worsening renal function.

Declarations

Ethics approval and consent to participate

This study analyzed deidentified MIMIC-IV version 3.1 and eICU Collaborative Research Database version 2.0 records, accessed by credentialed researchers through PhysioNet under the applicable data-use agreements [20,28]. The collection of patient information and creation of MIMIC-IV were reviewed by the Institutional Review Board at Beth Israel Deaconess Medical Center, which waived informed consent and approved sharing of that deidentified resource [20]. This secondary analysis involved no patient contact or new collection of identifiable private information.

Consent for publication

Not applicable. The manuscript contains no identifiable individual information.

Availability of data and materials

The MIMIC-IV version 3.1 and eICU Collaborative Research Database version 2.0 datasets are available to credentialed researchers through PhysioNet under their data-use agreements [20,28,29]. Individual-level derived data are not redistributed. Model equations and aggregate results are included in this article and Additional file 1. Analysis code, without restricted data, is provided in Additional file 2. Running the code requires separate authorized access to the source databases and preparation of the input files described in its README.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Jinhua Traditional Chinese Medicine Science and Technology Research Program (grant 2023KY05). The funder had no role in study design, data collection, analysis, interpretation, manuscript preparation, or the decision to submit.

Authors' contributions

SK contributed to conceptualization, methodology, formal analysis, visualization, and drafting. YP contributed to conceptualization, supervision, and critical revision. Both authors are accountable for the content of the manuscript.

Acknowledgements

We thank the patients whose deidentified data contributed to MIMIC-IV and eICU, and the clinicians, data scientists, and PhysioNet teams who created and maintain these research resources.

Additional file

Additional file 1 (.docx): Supplementary methods and Tables S1–S10. This file provides the complete-case model equations, predictor-imputation details, descriptive characteristics, missingness, internally evaluated performance, decision-curve values, the eICU feasibility gate, outcome availability, landmark timing, and reviewer-requested exploratory comparisons.

Additional file 2 (.zip): Analysis code and extraction queries used in the source-event reconciliation and model analyses. The archive contains no individual-level clinical data; its README describes the required authorized data access, original extraction queries, correction workflow, software environment, and execution limitations.

References

1. McCallum W, Tighiouart H, Testani JM, Griffin M, Konstam MA, Udelson JE, et al. Acute kidney function declines in the context of decongestion in acute decompensated heart failure. JACC Heart Fail. 2020;8:537–47. doi:10.1016/j.jchf.2020.03.009

2. Horiuchi Y, Wettersten N, van Veldhuisen DJ, Mueller C, Filippatos G, Nowak R, et al. Decongestion, kidney injury and prognosis in patients with acute heart failure. Int J Cardiol. 2022;354:29–37. doi:10.1016/j.ijcard.2022.02.026

3. Meekers E, Dauw J, Martens P, Dhont S, Verbrugge FH, Nijst P, et al. Renal function and decongestion with acetazolamide in acute decompensated heart failure: the ADVOR trial. Eur Heart J. 2023;44:3672–82. doi:10.1093/eurheartj/ehad557

4. Testani JM, Chen J, McCauley BD, Kimmel SE, Shannon RP. Potential effects of aggressive decongestion during the treatment of decompensated heart failure on renal function and survival. Circulation. 2010;122:265–72. doi:10.1161/CIRCULATIONAHA.109.933275

5. Rao VS, Ahmad T, Brisco-Bacik MA, Bonventre JV, Wilson FP, Siew ED, et al. Renal effects of intensive volume removal in heart failure patients with preexisting worsening renal function. Circ Heart Fail. 2019;12:e005552. doi:10.1161/CIRCHEARTFAILURE.118.005552

6. McCallum W, Tighiouart H, Testani JM, Griffin M, Konstam MA, Udelson JE, et al. Rates of reversal of volume overload in hospitalized acute heart failure: association with long-term kidney function. Am J Kidney Dis. 2022;80:65–78. doi:10.1053/j.ajkd.2021.09.026

7. Lorenzo M, Núñez G, Fuertes-Kenneally L, de la Espriella R, Villar S, Miró O, et al. Early glomerular filtration rates changes and risk of mortality in acute heart failure. The modifying role of admission renal function and decongestion. Eur J Intern Med. 2023;115:96–103. doi:10.1016/j.ejim.2023.05.037

8. Wattad M, Darawsha W, Solomonica A, Hijazi M, Kaplan M, Makhoul BF, et al. Interaction between worsening renal function and persistent congestion in acute decompensated heart failure. Am J Cardiol. 2015;115:932–7. doi:10.1016/j.amjcard.2015.01.019

9. Griffin M, Rao VS, Fleming J, Raghavendra P, Turner J, Mahoney D, et al. Effect on survival of concurrent hemoconcentration and increase in creatinine during treatment of acute decompensated heart failure. Am J Cardiol. 2019;124:1707–11. doi:10.1016/j.amjcard.2019.08.034

10. Metra M, Cotter G, Senger S, Edwards C, Cleland JG, Ponikowski P, et al. Prognostic significance of creatinine increases during an acute heart failure admission in patients with and without residual congestion. Circ Heart Fail. 2018;11:e004644. doi:10.1161/CIRCHEARTFAILURE.117.004644

11. Ruocco G, Nuti R, Giambelluca A, Evangelista I, De Vivo O, Daniello C, et al. The paradox of transient worsening renal function in patients with acute heart failure: the role of B-type natriuretic peptide and diuretic response. J Cardiovasc Med (Hagerstown). 2017;18:851–8. doi:10.2459/JCM.0000000000000546

12. Krishnamoorthy A, Greiner MA, Sharma PP, DeVore AD, Johnson KW, Fonarow GC, et al. Transient and persistent worsening renal function during hospitalization for acute heart failure. Am Heart J. 2014;168:891–900. doi:10.1016/j.ahj.2014.08.016

13. Kamio T, Ikegami M, Machida Y, Uemura T, Chino N, Iwagami M. Machine learning-based prognostic modeling of patients with acute heart failure receiving furosemide in intensive care units. Digit Health. 2023;9:20552076231194933. doi:10.1177/20552076231194933

14. Wang L, Zhou D, Xu P. A deep learning model for dynamic prediction of acute kidney injury in heart failure patients. Can J Cardiol. 2026;42:1605–15. doi:10.1016/j.cjca.2026.01.011

15. Peng X, Li L, Wang X, Zhang H. A machine learning-based prediction model for acute kidney injury in patients with congestive heart failure. Front Cardiovasc Med. 2022;9:842873. doi:10.3389/fcvm.2022.842873

16. Feng LH, Su T, Huang L, Liao T, Lu Y, Wu L. Development and validation of a dynamic nomogram for acute kidney injury prediction in ICU patients with acute heart failure. Front Med. 2025;12:1544024. doi:10.3389/fmed.2025.1544024

17. Li Y, Liu X, Wang X, Zhou L, Zhao C. Machine learning-based model for predicting acute kidney injury in patients hospitalized with heart failure: development and validation study. CJC Open. 2026;8:379–91. doi:10.1016/j.cjco.2025.11.019

18. Lee TH, Fan PC, Chen JJ, Wu VC, Lee CC, Yen CL, et al. A validation study comparing existing prediction models of acute kidney injury in patients with acute heart failure. Sci Rep. 2021;11:11213. doi:10.1038/s41598-021-90756-9

19. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10:1. doi:10.1038/s41597-022-01899-x

20. Johnson A, Bulgarelli L, Pollard T, Gow B, Moody B, Horng S, et al. MIMIC-IV (version 3.1). PhysioNet. 2024. https://doi.org/10.13026/kpb9-mt58

21. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378

22. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. doi:10.1136/bmj-2024-082505

23. Chawla LS, Bellomo R, Bihorac A, Goldstein SL, Siew ED, Bagshaw SM, et al. Acute kidney disease and renal recovery: consensus report of the Acute Disease Quality Initiative 16 Workgroup. Nat Rev Nephrol. 2017;13:241–57. doi:10.1038/nrneph.2017.2

24. Khwaja A. KDIGO clinical practice guidelines for acute kidney injury. Nephron Clin Pract. 2012;120:c179–84. doi:10.1159/000339789

25. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441. doi:10.1136/bmj.m441

26. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. 2016;74:167–76. doi:10.1016/j.jclinepi.2015.12.005

27. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data. 2018;5:180178. doi:10.1038/sdata.2018.178

28. Pollard T, Johnson A, Raffa J, Celi LA, Badawi O, Mark R. eICU Collaborative Research Database (version 2.0). PhysioNet. 2019. https://doi.org/10.13026/C2WM1R

29. Pollard T, Moody BE, Lehman L, Gow B, Fernandes C, Xie C, et al. PhysioNet as a global platform for biomedical research. Nature Health. 2026;1:792–5. doi:10.1038/s44360-026-00096-z

Tables

Table 1. Characteristics of the 12-hour landmark population

Characteristic	Overall	No persistent WRF	Persistent WRF	P value
Age, years	72.0 [62.0, 82.0] (n=433)	72.0 [62.0, 82.0] (n=370)	73.0 [59.5, 82.5] (n=63)	0.996
Female sex	183 (42.3%)	154 (41.6%)	29 (46.0%)	0.512
Race, overall	—	—	—	0.725
White race	301 (69.5%)	259 (70.0%)	42 (66.7%)	—
Black race	59 (13.6%)	49 (13.2%)	10 (15.9%)	—
Asian race	15 (3.5%)	14 (3.8%)	1 (1.6%)	—
Other/unknown race	58 (13.4%)	48 (13.0%)	10 (15.9%)	—
First-day SOFA score	4.0 [2.0, 6.0] (n=433)	4.0 [2.0, 5.0] (n=370)	5.0 [3.0, 8.0] (n=63)	<0.001
Diabetes	193 (44.6%)	160 (43.2%)	33 (52.4%)	0.177

Characteristic	Overall	No persistent WRF	Persistent WRF	P value
ICD-coded myocardial infarction	116 (26.8%)	96 (25.9%)	20 (31.7%)	0.337
Documented renal disease	193 (44.6%)	152 (41.1%)	41 (65.1%)	<0.001
Chronic pulmonary disease	199 (46.0%)	178 (48.1%)	21 (33.3%)	0.030
Cerebrovascular disease	28 (6.5%)	24 (6.5%)	4 (6.3%)	1.000
Peripheral vascular disease	67 (15.5%)	54 (14.6%)	13 (20.6%)	0.220
Malignancy	33 (7.6%)	26 (7.0%)	7 (11.1%)	0.300
Baseline serum creatinine, mg/dL	1.3 [1.0, 2.0] (n=433)	1.3 [1.0, 2.0] (n=370)	1.4 [1.2, 1.9] (n=63)	0.119
Creatinine change at landmark, mg/dL	0.0 [-0.1, 0.1] (n=393)	0.0 [-0.1, 0.1] (n=335)	0.2 [0.1, 0.5] (n=58)	<0.001
Cumulative urine output, mL	1425.0 [975.0, 2120.0] (n=433)	1503.5 [1031.2, 2278.7] (n=370)	1025.0 [582.5, 1427.5] (n=63)	<0.001
Pre-landmark binned MAP, mmHg	73.0 [66.0, 83.0] (n=402)	73.0 [65.5, 82.0] (n=347)	74.0 [66.5, 84.0] (n=55)	0.355
Cumulative furosemide dose, mg	80.0 [40.0, 160.0] (n=433)	80.0 [40.0, 157.9] (n=370)	80.0 [40.0, 200.0] (n=63)	0.104

MAP, mean arterial pressure; SOFA, Sequential Organ Failure Assessment; WRF, worsening renal function. Continuous values are median [Q1, Q3] with non-missing n; categorical values are n (%). Creatinine change uses the maximum value in the latest available pre-landmark six-hour bin minus baseline; binned MAP uses the first non-invasive value in its latest available pre-landmark bin. First-day SOFA describes illness severity during the first ICU day and was not a model predictor. ICD-coded myocardial infarction includes acute and prior infarction codes recorded for the same hospitalization. Continuous variables were compared with the two-sided Mann–Whitney U test. Binary variables were compared with Pearson's χ^2 test, or Fisher's exact test when any expected cell count was <5. Race was compared once across four categories using a Fisher–Freeman–Halton test with 200,000 Monte Carlo

resamples (fixed seed 20260930); its P value is approximate. Tests used available observations for each variable. No adjustment for multiple comparisons was made; these P values are descriptive and did not guide model specification.

SI conversion: multiply creatinine values in mg/dL by 88.4 to obtain µmol/L, and blood pressure values in mmHg by 0.1333224 to obtain kPa. The reported equations require the original units shown in each equation.

Table 2. Complete-case bootstrap internal validation

Landmark	Model	N/events	Metric	Corrected estimate [percentile range]
12 h	Base	368/52	AUROC	0.783 [0.721, 0.845]
12 h	Base	368/52	Brier score	0.100 [0.077, 0.122]
12 h	Base	368/52	Calibration intercept	0.035 [-0.454, 0.684]
12 h	Base	368/52	Calibration slope	1.020 [0.760, 1.364]
12 h	Base + dose	368/52	AUROC	0.791 [0.733, 0.852]
12 h	Base + dose	368/52	Brier score	0.101 [0.078, 0.123]
12 h	Base + dose	368/52	Calibration intercept	0.241 [-0.245, 0.834]
12 h	Base + dose	368/52	Calibration slope	1.148 [0.876, 1.465]
24 h	Base	357/51	AUROC	0.840 [0.791, 0.900]
24 h	Base	357/51	Brier score	0.092 [0.070, 0.113]
24 h	Base	357/51	Calibration intercept	0.085 [-0.521, 0.867]
24 h	Base	357/51	Calibration slope	1.050 [0.748, 1.474]
24 h	Base + dose	357/51	AUROC	0.844 [0.794, 0.898]
24 h	Base + dose	357/51	Brier score	0.093 [0.072, 0.114]
24 h	Base + dose	357/51	Calibration intercept	0.327 [-0.263, 1.107]

Landmark	Model	N/events	Metric	Corrected estimate [percentile range]
24 h	Base + dose	357/51	Calibration slope	1.202 [0.898, 1.605]

AUROC, area under the receiver-operating-characteristic curve. Ranges are empirical percentiles of corrected bootstrap replicate estimates; nominal 95% confidence-interval coverage was not established. Calibration intercept and slope were estimated jointly. Between-model differences in the text were calculated from unrounded estimates and may differ from subtracting the rounded values shown here.

Figures

Fig. 1. Study flow and external-evaluation gate. (a) The MIMIC-IV flow uses the earliest positive-dose, non-Rewritten furosemide event linked to the ICU stay, not the first dose of the hospitalization. Of 784 independent patients, 481 had a classifiable renal trajectory and 303 did not (227 without a pretreatment baseline creatinine; 3 without a post-treatment creatinine; 39 without sufficient follow-up to classify absence of an initial rise; 34 with an initial rise but insufficient 48–72-hour confirmation data). At 12 hours, 695 remained in the ICU, 262 lacked a classifiable outcome, and 433/63 entered the risk set; at 24 hours, the corresponding figures were 587, 204, and 383/57. The complete-case model populations were 368/52 and 357/51. (b) The eICU panel is a feasibility audit, not external validation; both landmarks failed the prespecified minimum of 40 future events. WRF, worsening renal function. The outcome assesses persistence of the first observed post-anchor creatinine rise; recurrent rises after recovery of the first episode are not reconsidered.

Alt text: Two-panel flow diagram. Panel a traces 1,364 acute-heart-failure intensive-care stays to 784 independent patients, 481 with classifiable renal outcomes, and 433 and 383 patients at the 12- and 24-hour landmarks. Panel b shows the eICU feasibility audit: 32 usable order-time records and only three future events at each landmark, below the 40-event threshold.

Fig. 2. Calibration after within-fold stochastic imputation. (a) The 12-hour landmark (433 patients/63 events). (b) The 24-hour landmark (383/57). Blue solid lines show the four-predictor base model; orange

dashed lines show the base model plus cumulative furosemide dose. Curves use pooled held-out predictions from 10-times repeated five-fold nested cross-validation, not the complete-case bootstrap estimates in Table 2. Points show predicted-risk decile observed proportions with Wilson intervals; shaded regions are patient-level conditional bootstrap percentile bands based on the stored predictions. The highest base-model predicted-risk decile contained 44 patients/24 events at 12 hours and 39/26 at 24 hours; its limited size contributes to uncertainty at high predicted risks. The gray dashed diagonal denotes ideal calibration.

Alt text: Two calibration panels for the 12-hour and 24-hour landmarks. Blue solid curves show the base model and orange dashed curves add furosemide dose. The full 0–1 predicted-risk scale is shown. The highest base-model risk decile has 44 patients at 12 hours and 39 at 24 hours, and the bands widen at high predicted risks.

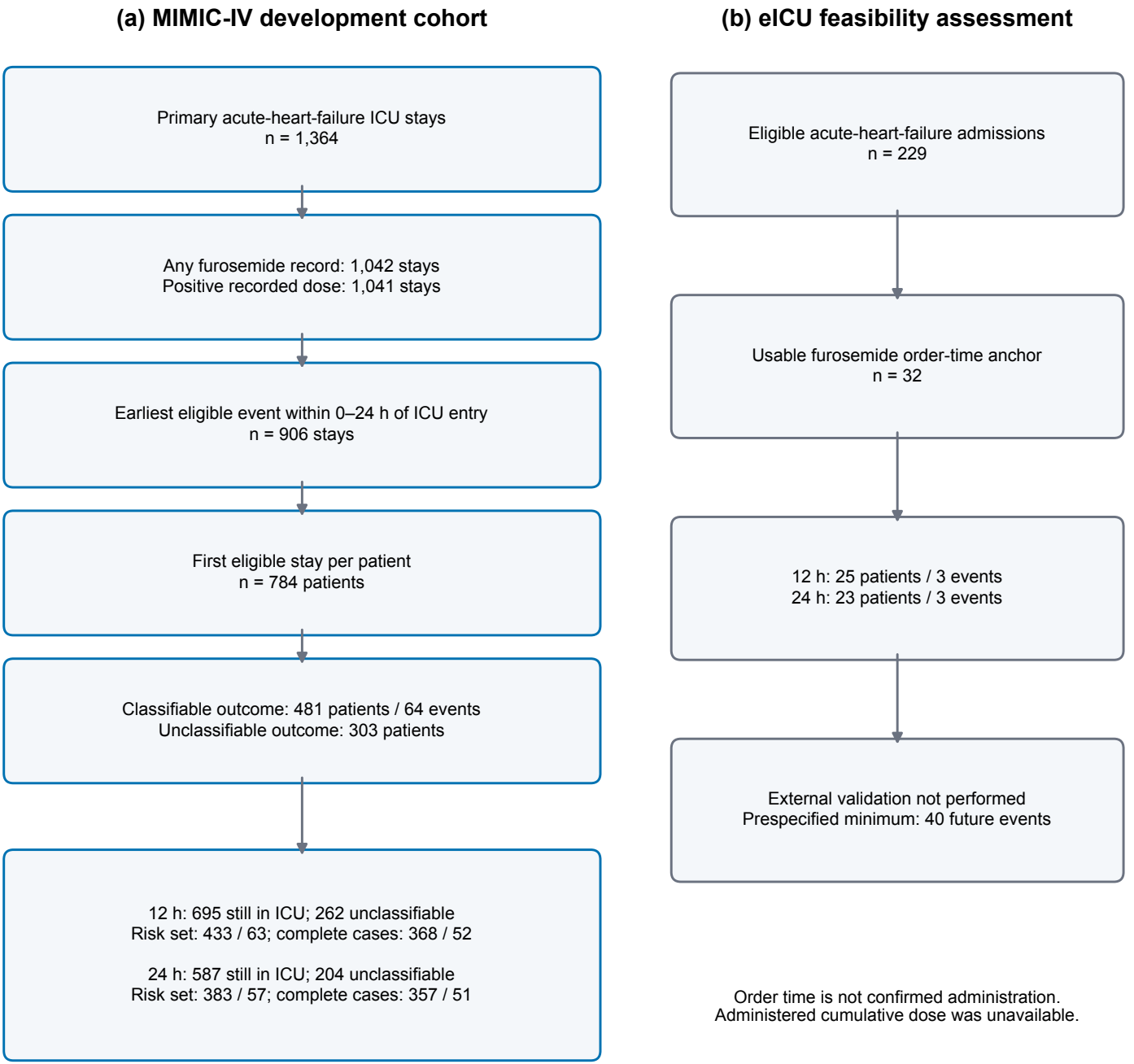
Fig. 3. Exploratory decision-curve analysis after within-fold stochastic imputation. (a) The 12-hour landmark (433 patients/63 events). (b) The 24-hour landmark (383/57). Blue solid lines show the four-predictor base model; orange dashed lines show the base model plus cumulative furosemide dose. The gray dashed curve and black horizontal line represent hypothetical intensified renal monitoring for all eligible patients and no extra monitoring beyond usual care, respectively. Curves show net benefit across thresholds from 2% to 30% using pooled held-out predictions from repeated nested cross-validation, not the complete-case estimates in Table 2. Shaded regions are patient-level conditional bootstrap percentile bands based on stored predictions. No actual intervention was linked to these risk thresholds; this internal calculation does not establish implementation benefit.

Alt text: Two decision-curve panels for 12-hour and 24-hour predictions. Across thresholds of 2% to 30%, the blue base-model and orange base-plus-dose point-estimate curves are above the treat-all and treat-none reference strategies. Shaded bands show internal bootstrap uncertainty; no clinical intervention was tested.

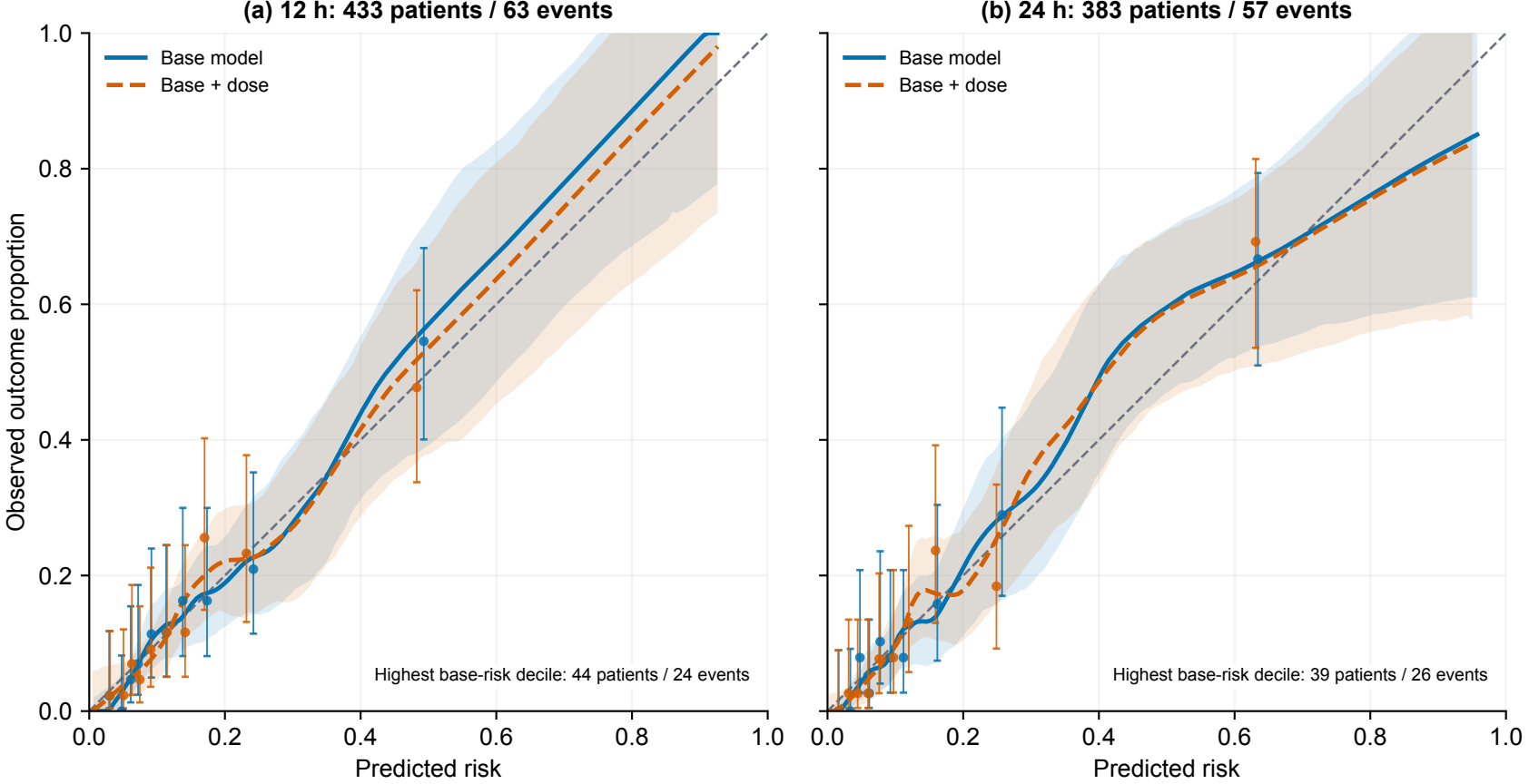
Fig. 4. Receiver-operating-characteristic curves from held-out predictions. (a) The 12-hour landmark (433 patients; 63 later-confirmed events). (b) The 24-hour landmark (383 patients; 57 later-confirmed events).

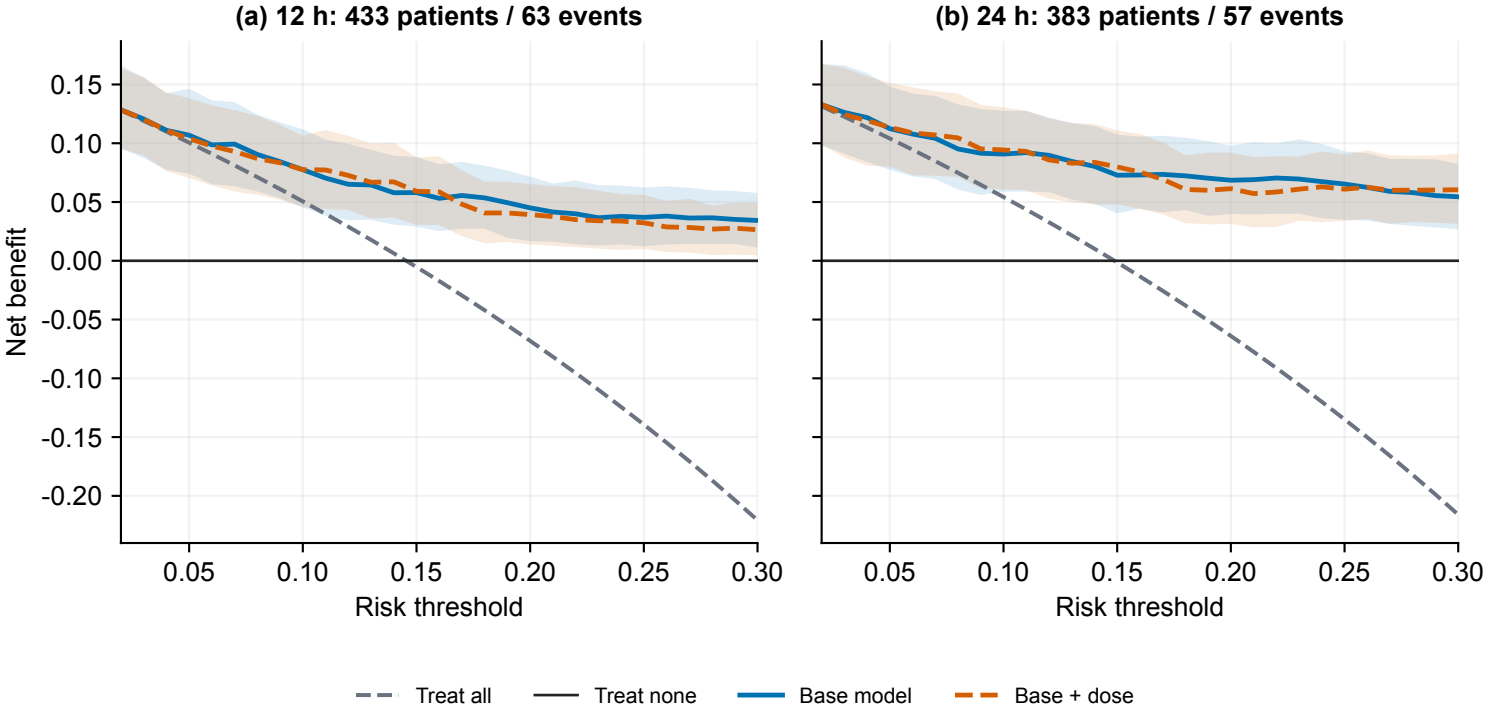
Blue solid lines show the four-predictor base model; orange dashed lines show the base model plus cumulative furosemide dose. Gray dashed diagonal lines indicate chance discrimination. Curves use pooled predictions from 10-times repeated five-fold nested cross-validation, with imputation fitted within each training fold. The AUROCs shown use the larger classifiable risk sets and are distinct from the complete-case, bootstrap optimism-corrected estimates in Table 2. AUC, area under the curve.

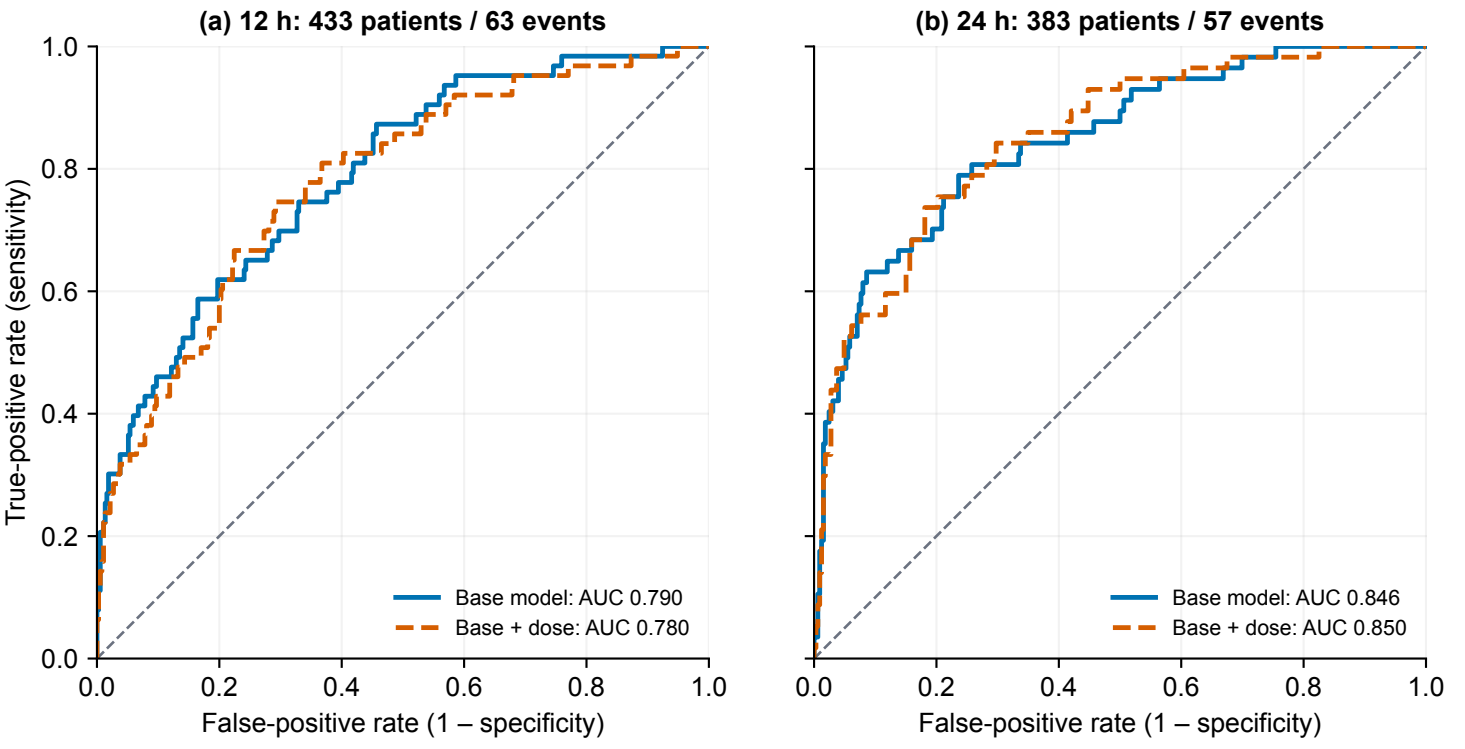
Alt text: Two receiver-operating-characteristic panels. At 12 hours, the base and base-plus-dose held-out areas under the curve are 0.790 and 0.780; at 24 hours they are 0.846 and 0.850. These curves come from classifiable risk sets after within-fold predictor imputation.



A single positive-dose, non-Rewritten event anchor was used for eligibility, predictors and follow-up.


(a) 12 h: 433 patients / 63 events
(b) 24 h: 383 patients / 57 events
Base model
Base + dose
Observed outcome proportion
Predicted risk
0.0
0.2
0.4
0.6
0.8
1.0
Highest base-risk decile: 44 patients / 24 events
Highest base-risk decile: 39 patients / 26 events


(a) 12 h: 433 patients / 63 events
(b) 24 h: 383 patients / 57 events
Net benefit
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
0.05
0.10
0.15
0.20
0.25
0.30
Risk threshold
Treat all
Treat none
Base model
Base + dose


(a) 12 h: 433 patients / 63 events
(b) 24 h: 383 patients / 57 events
True-positive rate (sensitivity)
False-positive rate (1 – specificity)
0.0
0.2
0.4
0.6
0.8
1.0
Base model: AUC 0.790
Base + dose: AUC 0.780
Base model: AUC 0.846
Base + dose: AUC 0.850

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1Revised.docx